Weekly report

1 Done

1.1 Reading

1.1.1 Randomization Techniques for Graphs（Kai Puolamaki）

- Goals:
- Perturb the original data and carry out the experiments with the randomized version.
- The randomization is controlled in such a way that the random datasets follow a certain distribution, which is typically chosen so that some properties (structural statistics) of the original data are maintained with a sufficient precision (**null hypothesis**).
- Preservations:
- The **number of nodes and edges**. (exactly)
- The **degree distribution**. (exactly)
- The **average clustering coefficient**. The clustering coefficient of a node v is the fraction of links the neighbors of the node have among them with respect to all possible such links. (approximately)

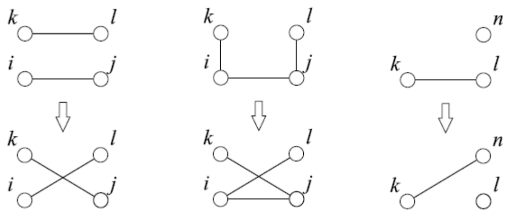
$$\mathrm{CC}(v) = \frac{|\{(u,w)|u,w \in \Gamma(v) \text{ and } (u,w) \in E\}|}{|\Gamma(v)|(|\Gamma(v)|-1)/2}$$

$$\mathrm{AvgCC}(G) = \frac{1}{|V(G)|} \sum_{v \in V(G)} \mathrm{CC}(v)$$

- The **characteristic path length** is the mean of all pairs shortest paths between the nodes. (approximately)

$$\mathrm{CPL}(G) = \frac{1}{|V(G)|^2} \sum_{u,v \in V(G)} d_G(u,v)$$

- Approaches:
- XSwap.
- Local-Swap.
- Flip.



(a) XSwap (b) Local-Swap (c) Flip

Input: Undirected graph G, distribution ρ and number of steps T.
`Randomize`(G, ρ, T)
$G_s \leftarrow G$
repeat
  Flip:
    Select an edge $(i,j) \in E(G_s)$
    Select (k,l) as either (i,j) or (j,i).
    Select node $n \in V(G_s)$
    if $k \neq n$ and $l \neq n$ and $(k,n) \notin E(G_s)$
      and $|\delta(k) - \delta(n)| = 1$ **then**
      $E(\widehat{G_s}) \leftarrow E(G_s) \backslash \{(k,l)\} \cup \{(k,n)\}$
      do with probability $\min(\rho(\widehat{G_s})/\rho(G_s), 1)$
        $G_s \leftarrow \widehat{G_s}$
      end do
  XSwap:
    Select edges $(i,j), (k,l) \in E(G_s)$
    if $(i,l) \notin E(G_s)$ and $(k,j) \notin E(G_s)$ **then**
      $E(\widehat{G_s}) \leftarrow E(G_s) \backslash \{(i,j),(k,l)\} \cup \{(i,l),(k,j)\}$
      do with probability $\min(\rho(\widehat{G_s})/\rho(G_s), 1)$
        $G_s \leftarrow \widehat{G_s}$
      end do
  LocalSwap:
    Select edges $(i,j), (i,k), (j,l) \in E(G_s)$
    if $k \neq l$ and $k \neq j$ and $l \neq i$ and
      $(i,l) \notin E(G_s)$ and $(j,k) \notin E(G_s)$ **then**
      $E(\widehat{G_s}) \leftarrow E(G_s) \backslash \{(i,k),(j,l)\} \cup \{(i,l),(j,k)\}$
      do with probability $\min(\rho(\widehat{G_s})/\rho(G_s), 1)$
        $G_s \leftarrow \widehat{G_s}$
      end do
until reaching T steps
return G_s
Output: Randomized graph G_s

1.1.2 **A Clustering Approach for Structural k-Anonymity in Social Networks using Genetic Algorithm** (Vikas Kumar Sihag)

- Definition:
  - **k-Anonymity for clustering**: at least k nodes are collapsed to form a supernode. Edge generalization is then employed to achieve indistinguishable nodes based on their relationships.
  - **SIL** (Structural Information Loss): the probability of error when trying to reconstruct the structure of the social network from the anonymized version, which is comprised of two components (intra-cluster, inter-cluster).
- Approach: genetic algorithm.

1.1.3 **Privacy Preserving Social Network Publication Against Mutual Friend Attacks** (Yan Fu)

- Definition:
  - The **NMF** of an edge: the number of mutual friends of the edge is the number of mutual friends of the two vertices.
  - **k-anonymous sequence**: a sequence vector is k-anonymous, if for any entry with value as v, there exist at least k - 1 other entries with value as v.
  - **k-NMF**: a graph is k-NMF anonymous if the number sequence of mutual friends of edges in is a k-anonymous sequence.
- Approach:
  - Algorithm ADD→Cleanup-operation.
  - BFS-based Edge Anonymization (BFSEA): candidates generation → candidates selection→candidates dynamic removal→edge anonymization.
  - Algorithm ADD&DEL
  - k_1-degree Anonymization Based on k_2-NMF:

1.1.4 **Line Graph or Scatter Plot? Automatic Selection of Methods for Visualizing Trends in Time Series** (Baoquan Chen)

This paper introduces an approach that automatically selects between a scatter plot and a line graph by comparing the density fields with the real trend density field. Combining automated approach with artificial identification, they evaluate how four elements exert effects on selections.

It is an interesting idea to take different band width into consideration. However, I think the encoding also plays an important role for users to identify the trends. Besides, I don't agree with the deletion of samples in the process of evaluation.

1.2 Coding for geo-privacy: render the trajectories and dwells needed for the new experiment.

1.3 Polishing our vast submission: test new datasets.